Electronic Version
Stylesheet Version v1.1.1

Description

[METHOD OF FABRICATING FLASH MEMORY CELL]

BACKGROUND OF INVENTION

[0001] Field of the Invention

[0002] The present invention relates to a semiconductor device and a method of manufacturing the same. More particularly, the present invention relates to a flash memory cell and a method of manufacturing the same.

[0003] Description of Related Art

[0004] Flash memory is a memory device that allows multiple data writing, reading, and erasing operations. In addition, the stored data will be retained even after power to the device is removed. With these advantages, it has been broadly applied in personal computer and electronic equipment.

[0005] A typical flash memory device has a floating gate and a control gate fabricated using doped polysilicon (the so-

called stack gate structure). The control gate is set above the floating gate with an inter-gate dielectric layer separating the two. Furthermore, a tunneling oxide layer is also set between the floating gate and an underlying substrate.

[0006] To write data into the flash memory, a bias voltage is applied to the control gate and the source/drain regions so that an electric field is generated to inject electrons into the floating gate. To read data from the flash memory, an operating voltage is applied to the control gate. Since the entrapment of charges inside the floating gate will directly affect the on/off status of the underlying channel, the on/off status of the channel can be construed as a data value of "1" or "0". Finally, to erase data from the flash memory, the relative potential between the substrate, the source region, the drain region or the control gate is raised. Hence, tunneling effect can be utilized to transfer electrons from the floating gate to the substrate or drain (source) via the tunneling oxide layer (the so-called substrate erase or drain (source) side erase) or from the floating gate to the control gate via the inter-gate dielectric layer.

[0007] However, it is difficult to control the flow of electrons

from the floating gate when data within a flash memory cell is erased. Thus, too much positive charge may be ejected from the floating gate leading to a phenomenon called over-erase. When over-erase is really significant, the channel underneath the floating gate may conduct even if no operating voltage is applied to the control gate. In other words, the channel permanently conducts leading to the production of incorrect data. To minimize over-erase phenomenon, many types of flash memories have a split gate design. Aside from a control gate and a floating gate, the split gate flash memory cell has a select gate (or an erase gate) above the substrate beside each sidewall of the control gate and the floating gate. The select gate is isolated from the control gate, the floating gate and the substrate through another inter-gate dielectric layer. When over-erase is significant, that is, the channel underneath the floating gate is conductive in the absence of an operating voltage applied to the control gate, the channel underneath the select gate remains off. With the select gate in a off mode, the drain region and the source region are non-conductive so that misreading of data is prevented.

[0008] However, because a split gate design demands a bigger

split gate area and a larger memory cell size, a split gate memory cell is more bulky than a convention stack gate memory cell. Ultimately, the level of integration cannot be further increased.

[0009] Furthermore, performance of the flash memory is closely related to the gate coupling ratio (GCR) between the floating gate and the control gate and the gate coupling ratio is dependent on the included area between the control gate and the floating gate. Therefore, if the included area between the control gate and the floating gate is large, the gate-coupling ratio is high resulting in a better device performance. Yet, increasing the included area between the control gate and the floating gate is more and more difficult when the level of integration is high.

[0010] Fig. 1 is a schematic cross-sectional view of a flash memory cell structure as disclosed in U.S. Patent No. 6,130,453. As shown in Fig. 1, the memory cell is formed within a trench in a silicon substrate 20 in order to increase the level of device integration. The memory cell includes two vertical floating gates 31a and 31b, a vertical bit line 32, spacers 25, a drain region 27, a source region 28, a silicon oxide cap layer 24 and a control gate (word line) 33.

[0011] Figs. 2A through 2D are schematic cross-sectional views showing the steps for fabricating a flash memory cell as disclosed in U.S. Patent No. 6,130,453. As shown in Fig. 2A, a substrate 20 having a thick patterned gate oxide layer 21 and a silicon nitride dielectric layer 23 that expose a trench 40 is provided. Thereafter, a thin gate oxide layer 22 is formed on the surface of the trench 30 and then polysilicon material is deposited into the trench 40 to form a polysilicon layer 31.

[0012] As shown in Fig. 2B, a source region 28 is formed in the substrate 20 on each side of the area reserved for forming a gate structure (that is, the trench 40). Thereafter, a reactive ion etching (RIE) operation is carried out to form a first floating gate 31a and a second floating gate 31b on the respective sidewalls of the trench 40. In the process, a trench 42 is also formed.

[0013] As shown in Fig. 2C, silicon nitride material is deposited on the surface of the trench 42 to form a silicon nitride layer 26. As shown in Fig. 2D, an oxidation process is carried out and then the silicon nitride layer 25 is etched to form silicon nitride spacers 25a. Thereafter, a drain region 27 is formed in the substrate 20 at the bottom of the trench 42. Finally, polysilicon material is deposited into

the trench 42 to form a polysilicon bit line 32.

[0014] According to the method disclosed in U.S. Patent No. 6,130,453, the spacers 25a between the floating gates 31a, 31b and the bit line 32 are formed by depositing dielectric material after forming the floating gates 31a, 31b but before forming the bit line 32. After forming the dielectric layer (the silicon nitride layer 25), an oxidation process is carried out followed by performing a reactive ion etching process to remove the dielectric layer at the bottom of the trench and expose the substrate 20.

[0015] However, in the process of removing the dielectric layer at the bottom of the trench, a portion of the dielectric layer attached to the sidewalls is also removed or etched. Hence, the performance of the memory cell is likely to be adversely affected.

SUMMARY OF INVENTION

[0016] Accordingly, the present invention is directed to a method of fabricating a flash memory cell capable of avoiding a deterioration of memory cell performance caused by a defective dielectric layer.

[0017] According to an embodiment of the invention, a method of fabricating a flash memory cell is provided. First, a substrate is provided. Thereafter, a patterned mask layer

is formed over the substrate. Using the patterned mask layer as an etching mask, the substrate is etched to form a trench. A first dielectric layer is formed over the substrate. A first gate and a second gate are formed on the respective sidewall of the trench. The first gate and the second gate are at a distance from each other. Furthermore, the first and the second gates expose a portion of the first dielectric layer at the bottom of the trench. A first source/drain region is then formed in the substrate at the bottom of the trench. Thereafter, a second dielectric layer is formed over the substrate and then a passivation layer is formed over the second dielectric layer. The passivation layer is formed using a semiconductor material or a conductive material. A portion of the passivation layer, the second dielectric layer and the first dielectric layer are removed to expose the substrate surface at the bottom of the trench. A third gate that completely fills the trench is formed. After removing the mask layer, a third dielectric layer is formed over the substrate. A fourth and a fifth gate are formed beside the respective sidewall of the first gate and the second gate. Finally, a second source/drain region is formed in the substrate on one side of the fourth and the fifth gate respectively.

[0018] In the present invention, an undoped polysilicon passivation layer is formed over the second dielectric layer. When a subsequent process (an etching process) for removing the material at the bottom of the trench to expose the substrate is carried out, the passivation layer is capable of protecting the second dielectric layer against possible damage. Hence, the flash memory cell can have an improved data retention capacity.

[0019] According to another embodiment of the present invention, an alternative method of fabricating a flash memory cell is provided. First, a substrate having a liner layer and a mask layer with an opening thereon and a trench in the substrate located within the opening is provided. Thereafter, a tunneling oxide layer is formed on the surface of the trench. A conductive layer fills the interior of the trench and the conductive layer is etched back to produce a conductive layer having a top section above the surface of the liner layer but lower than the surface of the mask layer. Thereafter, a pair of spacers is formed on the respective sidewall of the trench so that a portion of the conductive layer is covered. Using the spacers and the mask layer as an etching mask, a portion of the conductive layer is removed to form a first floating gate and a

second floating gate beside the respective sidewall of the trench. A first source/drain region is formed in the substrate at the bottom of the trench. Next, a first inter-gate dielectric layer is formed on the surface of the substrate and the trench and then a passivation layer is formed on the first inter-gate dielectric layer. The passivation layer is fabricated using a semiconductor material or a conductive material. Thereafter, a portion of the passivation layer, the first inter-gate dielectric layer and the tunneling oxide layer are removed to expose the substrate surface at the bottom of the trench. A control gate that completely fills the trench is formed. The top section of the control gate is at a level higher than the top section of the first floating gate and the second floating gate. After removing the liner layer and the mask layer, a second inter-gate dielectric layer is formed over the substrate. A first select gate and a second select gate are formed beside the respective sidewall of the spacers, the first floating gate and the second floating gate. Finally, a second source/drain region is formed in the substrate on one side of the first select gate and the second select gate respectively.

[0020] In an embodiment of the present invention, an undoped polysilicon passivation layer is formed over the first inter-

gate dielectric layer. When a subsequent process (an etching process) for removing the material at the bottom of the trench to expose the substrate is carried out, the passivation layer is capable of protecting the first inter-gate dielectric layer against possible damage. Hence, the flash memory cell can have an improved data retention capacity.

[0021] In an embodiment of the present invention, the first floating gate (or the second floating gate) and the control gate are formed in the trench within the substrate. Hence, size of the memory cell can be reduced and the level of integration can be increased. Furthermore, the included area between the first floating gate (or the second floating gate) and the control gate is related to the depth of the trench. Therefore, by forming a deeper trench, the included area between the first floating gate and the control gate is increased. With a larger included area, the gate-coupling ratio is increased leading to an increase in operating speed and performance of the device. Moreover, the length of the channel corresponds to the depth of the trench. Thus, abnormal punch-through between the first source/drain region and the second source/drain region can be avoided by forming a trench with a desirable

depth.

[0022] In addition, the method of fabricating the flash memory cell according to an embodiment of the present invention is capable of forming the common control gate that controls two memory cells at the same time. In other words, the first floating gate, the first select gate and the control gate together constitute a memory cell while the second floating gate, the second select gate and the control gate constitute another memory cell. Hence, the level of integration is increased and the production cost of the memory device is reduced.

[0023] It is to be understood that both the foregoing general description and the following detailed description are exemplary, and are intended to provide further explanation of the invention as claimed.

Brief Description of Drawings

[0024] The accompanying drawings are included to provide a further understanding of the invention, and are incorporated in and constitute a part of this specification. The following drawings illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention.

[0025] Fig. 1 is a schematic cross-sectional view of a conven-

tional flash memory cell structure.

[0026] Figs. 2A through 2D are schematic cross-sectional views showing the steps of fabricating a conventional flash memory cell.

[0027] Figs. 3A through 3J are schematic cross-sectional views showing the steps of fabricating a flash memory cell according to one embodiment of the present invention.

[0028] Figs. 4A through 4G are schematic cross-sectional views showing the steps of fabricating a flash memory cell according to another embodiment of the present invention.

DETAILED DESCRIPTION

[0029] Reference will now be made in detail to the present preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers are used in the drawings and the description to refer to the same or like parts.

[0030] Figs. 3A through 3J are schematic cross-sectional views showing the steps of fabricating a flash memory cell according to one embodiment of the present invention. As shown in Fig. 3A, a substrate 300 having at least a device isolation structure (not shown) thereon is provided. The device isolation structure has a linear layout for defining

an active region. The device isolation structure is formed, for example, by performing a local oxidation (LOCOS) or performing the well known process steps necessary for forming a shallow trench isolation (STI) structure.

[0031] A dielectric layer 301 is formed over the substrate 300. The dielectric layer 301 includes a silicon oxide layer formed, for example, by performing a chemical vapor deposition (CVD) process. In an alternative preferred embodiment, a thinner liner layer (not shown) is formed over the surface of the substrate by performing a thermal oxidation process. Thereafter, a patterned mask layer 304 is formed over the dielectric layer 301. The patterned mask layer 304 includes a silicon nitride layer formed, for example, by depositing a mask material layer in a chemical vapor deposition process and then performing a photolithographic/etching process. Using the patterned mask layer 304 as an etching mask, the substrate 300 is etched to form a trench 306 in the substrate 300.

[0032] Then, a first dielectric layer is formed over the substrate 300. The first dielectric layer is formed, for example, by forming a tunneling oxide layer 308 on the surface of the trench 306. The tunneling oxide layer 308 includes a silicon oxide layer formed by a thermal oxidation process,

for example. Thereafter, a conductive layer 310 is formed to fill the trench 306 by depositing doped polysilicon in a chemical vapor deposition process and then implanting ions into the undoped polysilicon layer.

[0033] As shown in Fig. 3B, the conductive layer 310 on the surface of the mask layer 304 is removed to form a conductive layer 310a in the trench 306. The conductive layer 310 above the mask layer 304 can be removed by etching back in a chemical-mechanical polishing operation, for example. Thereafter, a photolithographic and etching process (a patterning process) are carried out to form a trench 307 in the conductive layer 310a so that a pair of floating gates 314a and 314b are formed as shown in Fig. 3C. Afterwards, ions are implanted into the substrate 300 at the bottom of the trench 307 to form a first source/drain region 316 as shown in Fig. 3D.

[0034] As shown in Fig. 3E, a second dielectric layer is formed over the substrate 300. For example, an inter-gate dielectric layer 318 such as an oxide/nitride/oxide composite layer is formed over the mask layer 304 and the surface of the trench 307. Thereafter, a passivation layer 320 is formed over the inter-gate dielectric layer 318. The passivation layer 320 is formed of a semiconductor material or

a conductive material. For example, the passivation layer 320 is formed of polysilicon. The passivation layer 320 having a thickness of about 100Å is formed, for example, by performing a chemical vapor deposition.

[0035] As shown in Fig. 3F, a portion of the tunneling oxide layer 308, the inter-gate dielectric layer 318 and the passivation layer 320 are removed until a portion of the substrate 300 is exposed at the bottom of the trench 307 and forming an inter-gate dielectric layer 318a and a passivation layer 320a. The method of removing the tunneling oxide layer 308, the inter-gate dielectric layer 318 and the passivation layer 320 at the bottom of the trench 307 includes an anisotropic etching process (for example, a dry etching process). It should be noted that the inter-gate dielectric layer 318 disposed on the sidewalls of the trench 307 remains intact after the aforementioned process (the etching process) because the passivation layer 320 covers and protects the inter-gate dielectric layer 318.

[0036] As shown in Fig. 3G, a control gate 322 is formed within the trench 307 in the substrate 300. The control gate 322 is a layer of doped polysilicon material, for example.

[0037] As shown in Fig. 3H, the mask layer 304 is removed. It

should be noted that both the mask layer 304 and the liner layer are removed in the aforementioned step if a liner layer is formed over the substrate in Fig. 3A. Thereafter, a third dielectric layer such as an inter-gate dielectric layer 325 is formed over the substrate 300 covering the entire structure. The inter-gate dielectric layer 325 is a silicon oxide layer, for example.

[0038] Thereafter, a conductive layer 330 is formed over the inter-gate dielectric layer 325. In this embodiment, the conductive layer 330 is a doped polysilicon layer 326, for example. However, in an alternative embodiment, the conductive layer 330 is a composite layer including a doped polysilicon layer and a silicide layer.

[0039] As shown in Fig. 3I, spacers 332 are formed beside the respective sidewall of the floating gate 314a and the floating gate 314b.

[0040] As shown in Fig. 3J, using the spacers 332 as a self-aligned mask, a portion of the conductive layer 330 is removed so that select gates 334a and 334b are formed beside the respective sidewall of the floating gate 314a and 314b. The conductive layer 330 is removed, for example, by performing an anisotropic etching process. Finally, a second source/drain region 336 is formed in the substrate

300 beside the select gate 334a and the select gate 334b, thereby completing the fabrication of a pair of memory cells with a common control gate 322. The second source/drain region 336 is formed, for example, by implanting ions into the substrate 300.

[0041] In the aforementioned embodiment of the present invention, an undoped polysilicon passivation layer 320 is formed over the inter-gate dielectric layer 318. Hence, the passivation layer 320 can protect the inter-gate dielectric layer 318 from being damaged when an etching process is performed to remove the material at the bottom of the trench 307 to expose the substrate 300. With such protection, the flash memory cell can have a better data retention capability. Furthermore, the undoped polysilicon passivation layer 320 can serve as a buffer interface between the inter-gate dielectric layer 318 and the control gate 322. Moreover, the passivation layer 320 and the control gate 322 are formed by using the same material so that the two can be combined to form a single gate without additional processing steps.

[0042] Figs. 4A through 4G are schematic cross-sectional views showing the steps of fabricating a flash memory cell according to another embodiment of the present invention.

As shown in Fig. 4A, a substrate 400 having at least a device isolation structure (not shown) thereon is provided. The device isolation structure has a linear layout for defining an active region. The device isolation structure is formed, for example, by performing a local oxidation (LOCOS) or performing the steps necessary for forming a shallow trench isolation (STI) structure.

[0043] A liner layer 402 is formed over the surface of the substrate 400. The liner layer 402 includes a silicon oxide formed, for example, by performing a thermal oxidation process. In an alternative embodiment, a dielectric layer (not shown) is formed over the surface of the substrate 400 by performing a chemical vapor deposition process, for example. Thereafter, a mask layer 404 is formed over the liner layer 402. The mask layer 404 includes a silicon nitride layer formed, for example, by performing a chemical vapor deposition process. The mask layer 404, the liner layer 402 and the substrate 400 are patterned to form a trench 406 in the substrate 400.

[0044] A tunneling oxide layer 408 is formed on the surface of the trench 406. The tunneling oxide layer 408 is a silicon oxide layer formed, for example, by performing a thermal oxidation process. Thereafter, conductive material is de-

posited into the trench 406 and over the mask layer to form a conductive layer 410. The conductive layer 410 is fabricated using a doped polysilicon material. The conductive layer 410 is formed by forming an undoped polysilicon layer over the substrate 400 with a chemical vapor deposition process and then implanting ions into the undoped polysilicon layer.

[0045] As shown in Fig. 4B, an etching back process is carried out to remove a portion of the conductive layer 410. Hence, a conductive layer 410a is retained within the trench 406 such that the top section of the conductive layer 410a is at a level higher than the upper surface of the liner layer 402 but at a level lower than the upper surface of the mask layer 404. Thereafter, spacers 412 are formed on the respective sidewall of the trench 406 covering a portion of the upper surface of the conductive layer 410a. The spacers 412 are fabricated using a material having an etching selectivity different from the conductive layer 410a. The spacers 412 are formed, for example, by depositing insulating material to form an insulation layer (not shown) and then performing an anisotropic etching process to remove a portion of the insulation layer.

[0046] As shown in Fig. 4C, using the mask layer 404 and the spacers 412 as an etching mask, a portion of the conductive layer 410a is removed to form a first floating gate 414a and a second floating gate 4141b beside the respective sidewall of the trench 406. Thereafter, a first source/drain region 416 is formed in the substrate 400 at the bottom of the trench 406. The first source/drain region 416 is formed, for example, by implanting ions into the substrate 400.

[0047] As shown in Fig. 4D, an inter-gate dielectric layer 418 is formed over the substrate 400 and the surface of the trench 406. The inter-gate dielectric layer 418 is an oxide/nitride/oxide composite layer, for example. Thereafter, a passivation layer 420 is formed over the inter-gate dielectric layer 418. The passivation layer 420 is an undoped polysilicon layer having a thickness of about 100Å and is formed, for example, by a chemical vapor deposition process.

[0048] As shown in Fig. 4E, portions of the tunneling oxide layer 408, the inter-gate dielectric layer 418 and the passivation layer 420 are removed. Thus, a portion of the substrate 400 at the bottom of the trench 406 is exposed and an inter-gate dielectric layer 418a and a passivation layer

420a are formed. The method of removing portions of the tunneling oxide layer 408, the inter-gate dielectric layer 418 and the passivation layer 420 includes performing an anisotropic etching process (for example, a dry etching process). It should be noted that the inter-gate dielectric layer 418 disposed on the sidewalls of the trench 406 remains intact after the aforementioned process (the etching process) because the passivation layer 420 covers and protects the inter-gate dielectric layer 418.

[0049] A control gate 422 is formed within the trench 406. The top section of the control gate 422 is at a level higher than the top section of the floating gates 414a and 414b. The control gate 422 is fabricated using doped polysilicon material, for example. Thereafter, a cap layer 424 fills the trench 406 and covers the control gate 422.

[0050] As shown in Fig. 4F, the liner layer 402 and the mask layer 404 are removed. It should be noted that only the mask layer 404 must be removed if a thick dielectric layer instead of a liner layer 402 is formed over the substrate 400 in Fig. 4A. Thereafter, an inter-gate dielectric layer 425 is formed over the substrate 400 to cover the substrate 400 and the surface structures on the substrate 400. The inter-gate dielectric layer 425 is fabricated using silicon ox-

ide material, for example.

[0051] Thereafter, a conductive layer 430 is formed over the inter-gate dielectric layer 425. In this embodiment, the conductive layer 430 is a doped polysilicon layer 426 or a composite layer including a doped polysilicon layer 426 and a metal silicide layer 428, for example. After that, spacers 432 are formed beside the respective sidewall of the spacers 412, the first floating gate 414a and the second floating gate 414b.

[0052] As shown in Fig. 4G, using the spacers 432 as a self-aligned mask, a portion of the conductive layer 430 (the dope polysilicon layer 426 and the silicide layer 428) is removed to form a first select gate 434a and a second select gate 434b beside the respective sidewall of the spacers 412, the first floating gate 414a and the second floating gate 414b. The conductive layer 430 is removed by performing an anisotropic etching process, for example. Thereafter, a second source/drain region 436 is formed in the substrate 400 on one side of the first select gate 434a and the second gate 434b respectively so that a pair of memory cells having the same control gate 422 is produced. The second source/drain region 436 is formed, for example, by implanting ions into the substrate 400.

[0053] In an alternative embodiment, additional spacers (not shown) are formed beside the respective sidewall of the first select gate 434a and the second select gate 434b to facilitate the fabrication of a lightly doped drain (LDD) structure or a contact window.

[0054] In the aforementioned embodiment of the present invention, an undoped polysilicon passivation layer 420 is formed over the inter-gate dielectric layer 418. Hence, the passivation layer 420 is capable of protecting the inter-gate dielectric layer 418 against any damage when an etching process for removing the material at the bottom of the trench 406 to expose the substrate 400 is carried out. With such protection, the flash memory cell can have a better data retention capability. Furthermore, the undoped polysilicon passivation layer 420 can serve as a buffer interface between the inter-gate dielectric layer 418 and the control gate 422. Moreover, the passivation layer 420 and the control gate 422 are fabricated using an identical material so that the two can be combined to form a single gate without any need for further processing.

[0055] In the present invention, the floating gate and the control gate are formed in the trench within the substrate. Hence,

size of the memory cell can be reduced and the level of integration can be increased. Furthermore, the included area between the floating gate and the control gate is related to the depth of the trench. Therefore, by forming a deeper trench, the included area between the first floating gate and the control gate is increased. With a larger included area, the gate-coupling ratio is increased leading to an increase in operating speed and performance of the device. Moreover, length of the channel is also related to depth of the trench. Thus, abnormal punch-through between the first source/drain region and the second source/drain region can be avoided by forming a deeper trench.

[0056] In addition, the method of fabricating the flash memory cell according to the embodiment of the present invention is capable of forming a common control gate that controls two memory cells at the same time as shown in Fig. 4G. In other words, the first floating gate 414a, the first select gate 434a and the control gate 422 together constitute a memory cell while the second floating gate 414b, the second select gate 434b and the control gate 422 constitute another memory cell. Hence, the level of integration is increased and the production cost of the memory device is

reduced.

[0057] It will be apparent to those skilled in the art that various modifications and variations can be made to the structure of the present invention without departing from the scope or spirit of the invention. In view of the foregoing, it is intended that the present invention cover modifications and variations of this invention provided they fall within the scope of the following claims and their equivalents.